软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(11):3640−3656 [doi: 10.13328/j.cnki.jos.005828]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于对象位置线索的弱监督图像语义分割方法*

李 阳[1,2], 刘 扬[2], 刘国军[2], 郭茂祖[1,2,3]

[1](北京建筑大学 电气与信息工程学院,北京 100044)

[2](哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

[3](建筑大数据智能处理方法研究北京市重点实验室(北京建筑大学),北京 100044)

通讯作者: 刘扬, E-mail: yliu76@hit.edu.cn; 郭茂祖, E-mail: guomaozu@bucea.edu.cn

摘 要: 深度卷积神经网络使用像素级标注,在图像语义分割任务中取得了优异的分割性能.然而,获取像素级标注是一项耗时并且代价高的工作.为了解决这个问题,提出一种基于图像级标注的弱监督图像语义分割方法.该方法致力于使用图像级标注获取有效的伪像素标注来优化分割网络的参数.该方法分为 3 个步骤:(1) 首先,基于分类与分割共享的网络结构,通过空间类别得分(图像二维空间上像素点的类别得分)对网络特征层求导,获取具有类别信息的注意力图;(2) 采用逐次擦除法产生显著图,用于补充注意力图中缺失的对象位置信息;(3) 融合注意力图与显著图来生成伪像素标注并训练分割网络.在 PASCAL VOC 2012 分割数据集上的一系列对比实验,证明了该方法的有效性及其优秀的分割性能.

关键词: 图像语义分割;弱监督;深度卷积神经网络;注意力图;显著图

中图法分类号: TP391

中文引用格式: 李阳,刘扬,刘国军,郭茂祖.基于对象位置线索的弱监督图像语义分割方法.软件学报,2020,31(11):3640−3656. http://www.jos.org.cn/1000-9825/5828.htm

英文引用格式: Li Y, Liu Y, Liu GJ, Guo MZ. Weakly supervised image semantic segmentation method based on object location cues. Ruan Jian Xue Bao/Journal of Software, 2020,31(11):3640−3656 (in Chinese). http://www.jos.org.cn/1000-9825/5828.htm

Weakly Supervised Image Semantic Segmentation Method Based on Object Location Cues

LI Yang[1,2], LIU Yang[2], LIU Guo-Jun[2], GUO Mao-Zu[1,2,3]

[1](School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China)

[2](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

[3](Beijing Key Laboratory of Intelligent Processing for Building Big Data (Beijing University of Civil Engineering and Architecture), Beijing 100044, China)

Abstract: Deep convolutional neural networks have achieved excellent performance in image semantic segmentation with strong pixel-level annotations. However, pixel-level annotations are very expensive and time-consuming. To overcome this problem, this study proposes a new weakly supervised image semantic segmentation method with image-level annotations. The proposed method consists of three steps: (1) Based on the sharing network for classification and segmentation task, the class-specific attention map is obtained which is the derivative of the spatial class scores (the class scores of pixels in the two-dimensional image space) with respect to the network feature maps; (2) Saliency map is gotten by successive erasing method, which is used to supplement the object localization information missing by attention maps; (3) Attention map is combined with saliency map to generate pseudo pixel-level annotations and train the segmentation

* 基金项目: 国家自然科学基金(61671188, 61571164); 国家重点研发计划(2016YFC0901902)

Foundation item: National Natural Science Foundation of China (61671188, 61571164); National Key Research and Development Program of China (2016YFC0901902)

收稿时间: 2018-04-28; 修改时间: 2018-11-06; 采用时间: 2019-02-28; jos 在线出版时间: 2019-08-09

CNKI 网络优先出版: 2019-08-12 12:08:06, http://kns.cnki.net/kcms/detail/11.2560.TP.20190812.1207.006.html

network. A series of comparative experiments demonstrate the effectiveness and better segmentation performance of the proposed method on the challenging PASCAL VOC 2012 image segmentation dataset.

Key words: image semantic segmentation; weakly supervised; deep convolutional neural networks; attention map; saliency map

图像语义分割是指利用计算机的特征表达来模拟人类对图像的识别过程,为每个像素分配语义空间中的一个类别.其研究在场景理解、自动驾驶、机器人感知、气象预测、交通控制、人脸识别等领域具有广泛的应用价值.但是由于图像中对象的尺度、位置、光照、颜色等信息具有无穷多的变化形式,所以图像分割是计算机视觉领域极具挑战性的研究课题[1].

近年来,以卷积神经网络为代表的深度学习技术[2−7]的重大突破带来了图像语义分割性能的巨大提升[8−10].但是,此类方法的分割准确度很大程度上依赖于大量的像素级标注(pixel-level annotation)数据集[8−13].然而,收集这类数据集是一项昂贵和耗时的任务:平均需要耗时 4min 来标注一幅图像中的所有像素[14].此外,这也间接地反映出像素级标注是增强分割模型泛化能力的障碍.

为了克服这个问题,一些研究者尝试放宽图像标注的程度,提出了弱监督的语义分割方法[15−19].此类方法仅使用图像级标注(image-level annotation,明确地标注出图像中对象的类别),一方面,图像级标注的数据集更容易获得——平均仅需要 1s 来标注图像中对象的类别[20];另一方面,用于模型拓展的新类别图像集也更容易获取.因此,本文基于图像级标注提出了一种弱监督语义分割方法.

近两年,一些弱监督分割方法[16,21,22]通过引入有效的对象位置线索,使得其模型的分割性能得到显著地提升.这类方法的计算过程大体分为两个步骤:(1) 基于图像级标注获取对象的位置线索,构建伪像素标注(pseudo pixel-level annotation);(2) 利用伪像素标注训练深度卷积神经网络(deep convolutional neural network,简称 DCNN).所谓“伪”像素标注指的是并不是真实的准确的标注,但是它提供了对象在图像中的位置线索.可见,步骤(1)获取的伪像素标注将直接影响最后分割网络的性能.本文同样采用这个计算过程,主要关注如何通过图像级标注生成高质量的伪像素标注.

自顶向下策略在弱监督对象定位任务中发挥了很好的性能[23−26],因此,此类方法也被广泛地用于生成伪像素标注来指导弱监督语义分割任务.本文受 Simonyan 等人[25]启发,提出一种新的挖掘对象位置信息的方法,本文称这些具有类别信息的对象线索为注意力图(attention map,简称 AM).Simonyan 等人[25]通过计算类别得分对输入图像的导数获取注意力图,其结果并不理想,并且识别出的对象区域较为模糊.本文通过改进该方法,提出了分类与分割共享网络结构的注意力图获取方法.在同一网络结构上,通过计算空间类别得分对网络中间层特征的导数而生成注意力图,从而避免了网络的重复构建过程.本文(1) 采用空间类别得分(图像二维空间上像素的类别得分)对中间层特征求导,在很大程度上保存了对象的空间结构,使得识别出的对象更加完整;(2) 从目标类别注意力图中去除其他类别对象的噪声,生成更明确的目标类别对象位置信息,从而提高像素标注的准确性.

注意力图用于推理不同类别对象的位置信息,它挖掘出对于图像分类任务起关键作用的对象判别性区域.然而,虽然本文提出了空间类别得分的概念,但是由于网络中存在连续池化层,使得最后网络输出的尺度要远远小于图像的原始尺度,因此该方法还不足以检测出对象的全部区域,从而注意力图不足以作为伪像素标注训练分割网络.为了解决这个问题,我们借助于显著图检测模型[27,28],提出逐次擦除法来识别图像的前景对象.显著图与注意力图的区别是:(1) 注意力图上的对象具有语义类别信息;(2) 显著图上的前景对象是类别不可知的,它用于区分背景和前景信息.显著图和注意力图相互补充,并挖掘出对象的完整轮廓.最后,融合注意力图与显著图生成伪像素标注并训练分割网络.相比于其他弱监督图像语义分割方法,本文提出的方法有以下创新点:(1) 提出了一种分类与分割共享网络结构的注意力图获取方法,避免重复构建网络结构,并且该注意力图更具有判别性和准确性;(2) 提出了逐次擦除的显著图获取方法,使得模型在无需重复训练的基础上,能够检测出图像中存在的多个前景对象;(3) 通过融合注意力图与显著图生成高质量的伪像素标注,使得注意力图与显著图的信息相互补充,提供更精准的像素标注,从而提升分割网络的性能;(4) 采用了一个简单有效的计算框架,没有启发式的迭代训练挖掘的过程,从而提升了方法的可扩展性.

实验结果表明,本文提出的弱监督图像语义分割方法在 PASCAL VOC 2012 数据集合上表现出很好的性

能,与目前最先进的方法相比,取得了更好的分割准确率.

1 相关工作

图像语义分割根据其模型训练阶段所使用标注数据的监督程度,分为全监督(fully-supervised)、半监督(semi-supervised)和弱监督(weakly-supervised).近几年,语义分割的性能在深度卷积神经网络(DCNNs)的帮助下得到了显著的提升[8−10].全监督方法[29−32]在训练DCNNs时需要大量的像素标注,然而像素标注需要消耗大量的人力资源和时间.因此,半监督/弱监督语义分割问题受到了很多研究者的重视,并且提出了一些改进的方法.

半监督方法进一步弱化了数据标注的程度,采用 bounding box 标注、点标注、少量像素标注等.Lin 等人[33]使用 scribbles 标注(提供对象上少量像素的标签)来训练分割网络.Bearman 等人[14]融合了点标注与对象先验信息.Dai 等人[34]使用 bounding box 标注迭代生成对象候选集和训练卷积网络.Papandreou 等人[35]借助少量像素标注提升分割网络的性能.

弱监督是在半监督的基础上再进一步降低数据标注的成本,在仅有图像级标注的情况下训练分割模型.一些早期的工作[36,37]将弱标记语义分割看作为多示例学习(multiple instance learning)问题,即如果图像中至少有一个像素是正例,那么该图像被看作为正例;如果全部像素都是负例,那该图像也被看作为负例.此外,Pathak 等人[19]在损失函数中加入一些约束项,将分割问题看作约束优化问题.Papandreou 等人[35]采用期望最大化(expectation- maximization)方法交替预测像素类别和优化 DCNNs 参数.Hong 等人[38]利用图像数据集之间的知识迁移性引导目标数据集的分割网络优化过程.Wei 等人[39]提出了两个网络的训练策略.然而由于缺乏有效的对象位置信息,上述方法的分割性能还有很大的提升空间.

目前,一些分割方法[16,22,40]通过引入对象位置信息来生成伪像素标注并训练分割网络,其分割性能得到显著提升.可见:生成的伪像素标注的质量将直接影响分割网络的训练过程,从而影响最后的分割结果.生成伪像素标注的策略可以分为两类:图像挖掘和区域挖掘.其中,图像挖掘策略侧重图像的整体性,它假设简单的图像(只有一个类别对象,对象位于图像的中心区域,背景简单)的像素标注可以通过显著图检测[41]和共分割(co-segmentation)[42]获取.然后利用这些简单图像初步训练分割网络,并预测复杂图像的像素标注.这类方法通常需要大量额外图像数据,从而增加了数据获取难度.另一类区域挖掘方法[16,22,43]通过分类网络生成对于分类任务具有关键作用的判别性区域,目前被广泛用于生成像素标注的区域挖掘方法[23−26]基本采用自顶向下的技术.Zhou 等人[23]将分类网络中的全连接层替换为全卷积层(fully convolutional layer)和全局均值池化层(global average pooling layer),根据分类损失函数训练网络参数,并获取每个类别的激活图(class activation map,简称 CAM).Zhang 等人[24]提出一种新的反向传递法(excitation back propagation),通过网络的反向传递过程识别每个类别的判别性区域.Simonyan 等人[25]同样采用网络反向传递过程,计算类别得分对输入图像的导数,从而获取判别性的区域信息.Selvaraju 等人[26]提出泛化的 CAM 模型.

上述方法中,CAM 模型是应用最广泛的获取对象位置信息的方法[16,22,43,44].但是该方法只能识别出对象中最具判别性的区域,而非完整的对象.为了提高伪像素标注的准确性,Kolesnikov 等人[16]通过全局加权 rank-pooling 操作扩展判别性的区域.Wei 等人[22]采用对抗擦除的方法,迭代擦除当前最具有判别性的区域,并重新训练分类网络,最后合并每次擦除的区域用于生成伪像素标注.Kim 等人[44]通过两阶段法挖掘出对象的位置信息.此外,Shimoda 等人[45]通过改进 Simonyan 等人的方法[25]获取每类对象的判别性区域,并通过条件随机场优化分割结果.本文将上述各种方法获取的对象判别性区域统称为注意力图.

本文方法同样致力于获取高质量的伪像素标注.受 Simonyan 等人[25]的启发,本文提出新的获取对象判别性区域的方法,称为注意力图(attention map,简称 AM).与其他方法[22,25,45]需要使用不同的网络来获取注意力图和完成分割任务相比,我们使用共享的网络结构,端对端(end-to-end)地为获取注意力图和分割结果训练该网络.与 Hou 等人的方法[46]类似,本文同样采用显著图(saliency map,简称 SM)检测(挖掘图像中的前景对象,不具有语义类别信息)来辅助分割过程.Hou 等人[46]只考虑简单的图像(图像只包含单一类别对象),而本文所提出的逐次擦除法能够挖掘出图像中的多个类别对象.与 Wei 等人的方法[22]不同,本文无需擦除检测到的前景对象之后重复

训练分类网络.最后,本文通过融合注意力图与显著图生成伪像素标注,并且训练分割网络.

2 弱监督图像语义分割

本文方法的流程如图 1 所示.在生成注意力图之前,首先训练图像分类网络并且保存该网络的参数.给定图像 X 和其相对应的图像级标注 Z(图 1 中,Z={人,摩托车}),首先利用分类网络的反向传递过程产生各个语义类别 $c\in Z$ 的注意力图(第 2.2 节).之后,为了获取对象更加准确的位置信息,通过逐次擦除法(第 2.3 节)生成该图像的显著图(无类别信息),并与注意力图融合,得到该图像的伪像素标注(第 2.4 节).最后,利用伪像素标注计算分割网络的损失函数(公式(1)),通过随机梯度下降算法(stochastic gradient descent,简称 SGD)优化分割网络的参数.

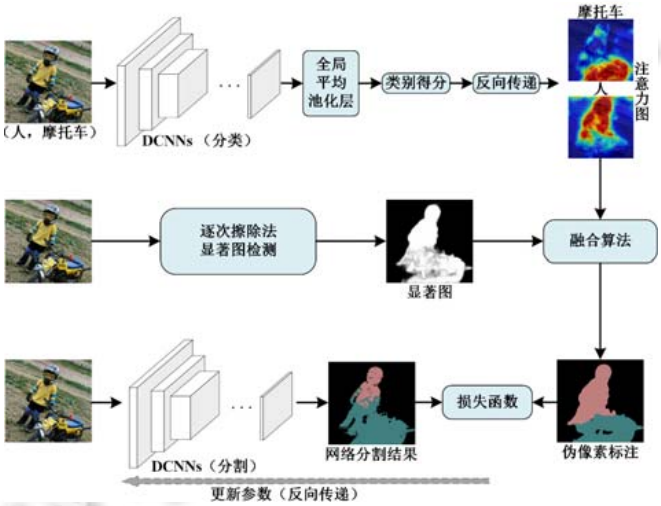

Fig.1 Pipeline for weakly supervised image semantic segmentation

图 1 弱监督图像语义分割流程

2.1 问题描述

弱监督语义分割采用弱标注的图像样本集合 I.对于每一幅图像 $X\in I$,分割图 $Y=\{y_1,y_2,\ldots,y_n\}$ 是 n 个像素位置上的语义标签集合.其中,y_i 从语义标签 $C=\{c_0\}\cup C'$ 中选择一个语义类别,c_0 表示为背景标签,$C'=\{c_1,\ldots,c_{CT}\}$ 表示为所有前景类别标签集合(CT 为图像集包含的前景类别的个数).本文将具有 N 幅图像的训练集表示为 $D=\{X_i,Z_i\}_{i=1}^{N}$,其中,$Z_i\subset C'$ 为图像级(image-level)标注,用于编码图像中存在的前景类别.模型的目标是利用训练集 D 来优化深度卷积神经网络 $f(X;\theta)$ 的参数 θ,该网络为像素 $m(m\in\{1,2,\ldots,n\})$ 分配标签 $c\in C$ 的条件概率建模,即 $f_{m,c}(X;\theta)=P(y_m=c|X;\theta)$.为了简洁表述,本文省略掉图像索引号,将 X_i 简化为 X,Z_i 简化为 Z.在有监督的情况下,深度卷积神经网络的目标函数定义为

$$Loss(\theta)=\log P(Y\mid X;\theta)=\sum_{m=1}^{n}\log P(y_m\mid X;\theta) \tag{1}$$

如果能够获取像素级标签,就可以直接计算每个像素真实语义类别的概率值,并且通过 mini-batch 随机梯度下降算法优化网络参数 θ.但是当只有图像级标签时,每个像素的真实语义类别是不可知的.那么需要通过图像的标注 Z 来获取一些对象的位置线索,进而构建伪像素标注.

本文通过 3 个过程获取伪像素标注(如图 1 所示,其中,分类网络与分割网络是同一网络结构):首先,计算图像空间类别得分对卷积神经网络特征层的导数,并生成具有类别信息的注意力图;然后,采用逐次擦除法获取图像前景对象的显著图;最后,通过融合注意力图与显著图生成伪像素标注来优化分割网络的参数.

2.2 生成注意力图

Simonyan 等人[25]利用类别得分对图像的导数来获取注意力图,该方法表明,可以通过网络反向传递过程来

定位对象的位置.受该方法的启发,本文提出了具有空间类别得分概念和分类/分割共享网络结构的注意力图生成方法,从而避免了额外构建分类网络的过程.此外,考虑到 Simonyan 等人的方法存在梯度减弱或者消失等问题,本次采用空间类别得分对网络中间特征层的导数值作为注意力图,从而可以有效地保留更多的高层语义并获取更加准确的对象位置信息.

为了获取图像中对象的注意力图,我们在分割网络最后的卷积层之后再加入一个卷积层(卷积核尺寸为1×1×|C'|,对于 PASCAL VOC 数据集,|C'|=20)和一个全局平均池化层,用于获取图像的|C'|个类别得分.正如 Oqual 等人[47]将多标签分类问题看作|C'|个独立的二分类问题来训练网络参数,那么当前图像分类任务的损失函数为

$$Loss_c(\theta)=-\frac{1}{|Z|}\sum_{c\in Z}\log P(c\mid X)-\frac{1}{|\bar{Z}|}\sum_{c\in\bar{Z}}\log(1-P(c\mid X)) \tag{2}$$

其中,Z 为图像真实的语义类别集,$\bar{Z}=C'\setminus Z$ 为图像中缺失的语义类别集,$P(c|X)$为网络预测的类别概率.根据公式(2)训练卷积神经网络,就可以得到每幅图像的类别得分,同时还可以得到|C'|个类别的对象空间位置得分 H.该空间类别得分 H 包含了对象在图像的位置信息,它虽然不能提供精准的对象定位线索,但是它可以在二维空间范围内给出对象在不同位置的类别得分.因此,不同于 Simonyan 等人的方法[25]使用图像类别得分对图像求导数,本文通过阈值化的 H 对中间卷积层的特征求导数.语义类别 c 的空间类别得分 H_c 尺度为 $w\times h$,根据设定好的阈值(实验中,$threshold$=0.8):如果 $H_c^{i,j}>threshold(i\in\{1,2,...,w\},\ j\in\{1,2,...,h\})$,则 $H_c^{i,j}$=1;否则 $H_c^{i,j}$=0.那么 H_c 对第 n 层特征 L_n 在激活值 L_n^0 的导数为

$$V_c^n=\left.\frac{\partial H_c}{\partial L_n}\right|_{L_n^0} \tag{3}$$

其中,V_c^n 通过网络反向传递过程计算得到.因为网络的最大池化(max pooling)运算使得 V_c^n 的尺度要小于原始图像的尺度,所以下一步通过上采样方法将 V_c^n 恢复为原始图像的尺度($W\times H$),记为 W_c^n.那么第 n 层的注意力图 $\tilde{A}_c^n\in R^{W\times H}$,$\tilde{A}_c^{n,i,j}=\max_{k_n}|W_c^{k_i,i,j}|$.其中,$k_n$ 表示第 n 层特征的通道(channel)个数,k_i 表示第 n 层特征的第 k_i 个通道,i,j 分别为二维空间的坐标.最后,将注意力图 $\tilde{A}_c^n$ 归一化.

对于图像的类别标签 Z 中的每个语义类别 c,上述方法都可以生成第 n 层的注意力图 $\tilde{A}_c^n$.为了更准确地区分每个类别的注意力图,并且解决不同类别的注意力图相互重叠的问题,我们借鉴了 DCSM 方法[45],将当前类别 c 的注意力图 $\tilde{A}_c^n$ 减去其他语义类别 $Z\backslash c$ 的注意力图,从而去掉其他语义类别的噪声:

$$\tilde{A}_c^{n,i,j}=\sum_{c'\in Z}\max(\tilde{A}_c^{n,i,j}-\tilde{A}_{c'}^{n,i,j},0)\,[c\neq c'] \tag{4}$$

最后整合各个层的注意力图,生成类别 c 的注意力图 $A_c^{i,j}$:

$$A_c^{i,j}=\frac{1}{|L|}\sum_{n\in L}\tanh(\tilde{A}_c^{n,i,j}) \tag{5}$$

其中,L 为上述方法所采用的网络特征层集合.图 2 描述了注意力图的生成过程.

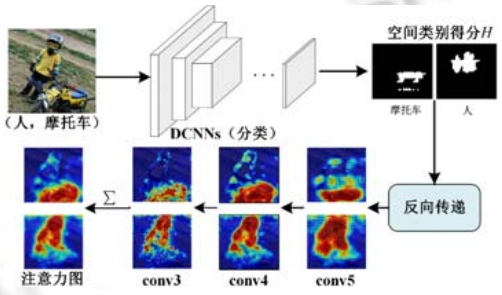

Fig.2 Procedure of obtaining attention map (AM)

图 2 注意力图生成过程

2.3 生成显著图

注意力图检测出对于分类任务最具有判别性的区域而非物体的全部轮廓(如图 2 所示).因此为了获取更加准确的对象位置信息,本文提出逐次擦除的显著图获取方法,目的是获取图像中前景对象的位置(该前景对象不具有语义类别信息).一些显著对象检测算法[27,28]的主要问题是,他们不能检测出图像中的多个对象.为了解决这个问题,本文提出了逐次擦除法来尽可能地识别出图像中存在的所有对象.其具体过程在算法 1 中给出了详细的描述.

算法 1. 逐次擦除算法.

输入:图像 X,阈值 $threshold_t(t\in\{1,2,3\})$,颜色均值 RGB_{mean}.

输出:显著图 S.

1: Initialize $S=zeros(W\times H)$ //其中,$W\times H$ 为图像 X 的尺度

2: **For** erasing number $t\in\{1,2,3\}$ **do**

3: Putting image X into saliency network SD;

4: Obtaining saliency map S_t

5: **If** t=1

6: $S=S_t$

7: **Else For each** pixel j

8: $S^j=\max(S^j,S_t^j)$

9: **End If**

10: $X^e=X(Index(S_t\geqslant threshold_t))=RGB_{mean}$

//擦除图像 X 的显著区域,其中 $Index(S_t\geqslant threshold_t)$表示 S_t 中大于等于 $threshold_t$ 的像素位置索引值

11: $X=X^e$

12: **End For**

如图 3 所示,逐次擦除法可以检测出更多的对象.与 AE(adversarial erasing)方法[22]相比,逐次擦除除法在不需要重复训练网络的情况下,能够获取图像中更多的对象位置信息.

在实验中,我们采用了 WSS(weakly supervised saliency)[27],DHSN(deep hierarchical saliency network)[28]两个方法作为基础显著对象检测模型 SD(saliency detector).WSS 方法基于图像级标注训练显著网络,DHSN 方法使用不具有语义类别信息的前景标注训练其显著图检测网络.

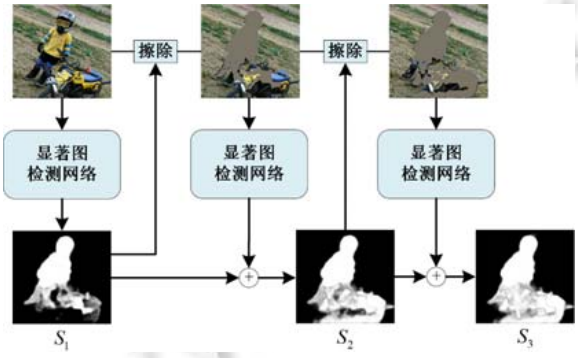

Fig.3 Procedure of the successive erasing method

图 3 逐次擦除法的过程

2.4 融合注意力图与显著图生成伪像素标注

前面已经提到,注意力图可以为每个语义类别提供该类别对象的位置信息,但是这类信息只能标注出每类对象最具有判别性的区域;显著图标注图像的前景对象,但是不具有任何类别信息.这两种线索都不足以单独作

为伪像素标注来训练分割网络,因此我们提出了融合算法(算法 2).对于每幅图像,该算法首先计算每个类别注意力图 A_c 与显著图 S 的均值 M_c,这样可以补充注意力图中没有被检测到的对象区域;之后,通过设定背景阈值 T_{c_0} (实验中,$T_{c_0}=0.2$)来限定一部分像素被标注为背景 M_{c_0};最后根据 M_c 与 M_{c_0} 获取伪像素标注 M.

算法 2. 融合算法.

输入:图像 X 的类别标签 Z;注意力图集合 $A=\{A_c\},\forall c\in Z$;显著图 S;背景阈值 T_{c_0}.

输出:像素标注 M.

1: Initialize $M=zeros(W\times H)$ //其中,$W\times H$ 为图像 X 的尺度
2: Initialize $M_{fg}=zeros(W\times H,|Z|)$, $M_{c_0}=zeros(W\times H)$
3: **For each** semantic label $c\in Z$ **do**
4: **For each** pixel p in image X
5: $M_{fg}(p,c)=mean(A_c^p,S^p)$
6: **End for**
7: **End for**
8: **For each** pixel p in image X **do**
9: $M_{\max}(p)=\max(M_{fg}(p,c),axis=2)$
10: **End for**
11: **For each** pixel p in image X **do**
12: **If** $M_{\max}(p)<T_{c_0}$ **then** $M_{c_0}(p)=1$ //标注背景
13: **End if**
14: **End for**
15: Concatenate $M_{all}=[M_{fg},M_{c_0}]$ //合并背景与前景
16: **For each** pixel p in image X **do**
17: $M(p)=\text{argmax}M_{all}(p)$
18: **End for**

最后,利用伪像素标签 M 作为像素标注并训练分割网络,其损失函数为公式(1).分割网络去掉了用于获取注意力图的最后卷积层和全局平均池化层,在倒数第 2 个卷积层之后加入像素分类层(softmax 层).

2.5 全连接条件随机场优化分割结果

由于分割网络中存在多层池化运算,因此最终的分割输出尺度远小于原始图像的尺度,以至于不能很好地分割出对象的轮廓.为了解决这个问题,我们采用全连接条件随机场模型(dense conditional random fields,简称 dense CRF)[48]优化分割结果.图像 X 的每个像素被当作一个节点,每个节点与其他节点之间是相互连接的,那么像素类别 Y 的能量函数为

$$E(Y)=\sum_n \phi(y_n)+\sum_{(n,n')}\psi(y_n,y_{n'}) \tag{6}$$

其中,$\phi(y_n)$表示为第 n 个像素分配语义类别的惩罚项;$\psi(y_n,y_{n'})$为平滑函数,用于惩罚相邻两个像素的语义标签是否一致.本文定义$\phi(y_n=c)=-\log(P(y_n=c|X;\theta))$,即通过分割网络计算得到的第 n 个像素被分配为类别 $c(c\in C)$的概率.

基于 dense CRF 模型[48],$\psi(y_n,y_{n'})=\mu(y_n,y_{n'})k(f_n,f_{n'})$.其中,$\mu(y_n,y_{n'})$表示相邻像素之间标注的兼容性,定义为

$$\mu(y_n,y_{n'})=\begin{cases}1, & \text{if } y_n\neq y_{n'}\\ 0, & \text{otherwise}\end{cases} \tag{7}$$

此外,$k(f_n,f_{n'})$是高斯核函数,f_n 与 $f_{n'}$分别为像素 n 及 n'的特征,表示为

$$k(f_n,f_{n'})=w_1\exp\left(-\frac{|S_n-S_{n'}|}{2\gamma_\alpha^2}-\frac{|T_n-T_{n'}|}{2\gamma_\beta^2}\right)+w_2\exp\left(-\frac{|S_n-S_{n'}|}{2\gamma_\delta^2}\right) \tag{8}$$

其中,$w_1,\gamma_\alpha,\gamma_\beta,w_2,\gamma_\delta$为模型参数.根据 dense CRF 参数推理方法[48],$w_1,w_2,\gamma_\alpha,\gamma_\beta,\gamma_\delta$通过网格搜索技术(grid search)优化(实验中,$w_1=w_2=1,\gamma_\alpha=30,\gamma_\beta=10,\gamma_\delta=3$).$k(f_n,f_{n'})$的第 1 个核函数依赖于像素位置(表示为 S_n)及颜色特征(表示为 T_n),而第 2 个核函数仅依赖于像素的位置特征.Krahenbuhl 等人[48]提供了上述能量函数(公式(6)))的近似推理.最后,经过全连接条件随机场模型的优化,分割结果可以更好地匹配对象的轮廓.

3 实验结果及分析

为了验证本文所提方法能够获得更好的分割结果,本节在 PASCAL VOC 数据集上进行了一系列验证和对比实验.第 3.1 节详细地描述了实验的各种设置.第 3.2 节~第 3.5 节列出本文所提方法的分割结果.

3.1 实验设置

- 数据集

为了验证本文所提方法的有效性,我们采用 PASCAL VOC 2012[49]分割数据集合,包括 20 个前景对象类别和一个背景类别.其原始分割数据集中有 1 464 幅训练图像、1 449 幅验证图像及 1 456 幅测试图像.遵循现有方法的惯例[16,22],本文拓展了训练数据集[50]:10 582 幅图像.与其他方法的对比实验分别在验证集及测试集上进行.本文方法的实验结果均是通过官方 PASCAL VOC 提供的评估服务器获取.

- 评价指标

本文使用语义分割标准度量——平均 IoU(mean intersection over union)来衡量分割效果.每幅图像的 IoU 定义为

$$IoU=\frac{GT\cap PS}{GT\cup PS} \tag{9}$$

其中,GT 为图像的真实分割,PS 为图像的预测分割.实验中计算 21 个类别的平均 IoU 值.

- 网络结构

(1) 显著图网络

本文使用 WSS[27]与 DHSN[28]作为显著图检测器,并通过逐次擦除法来发现图像中存在的多个对象.这两个模型将 VGG-16[4]作为基础网络结构.在逐次擦除法中,本文设定显著得分大于 0.7 的像素区域被擦除.

(2) 注意力图获取网络与分割网络

本文基于 VGG_16_LargeFOV(large field of view)[9]模型构造注意力图获取网络与分割网络.为了获取注意力图,本文在分割网络的最后卷积层外追加了一个卷积层(输出为 20 个通道,卷积核大小为 1×1×20).除最后两个卷积层外,我们还采用在 ImageNet[3]数据集上预训练好的参数来初始化模型.最后两层参数由正太分布($N(0,0.01)$)随机初始化.输入图像被随机切分为 321×321 尺度,最后网络输出 21 个 41×41 尺度的分割图.

- 网络训练

为了获取注意力图,本文首先用 PASCAL VOC 2012 训练集,基于分类损失函数(公式(2))与 mini-batch 随机梯度下降(SGD)算法训练上述网络.初始学习率为 0.001,每经过 2 000 次迭代,学习率降低 10 倍.此外,设定 dropout 层的 $drop\_rate$=0.5,动量 $momentum$=0.9,权值衰减率 0.0005.对于分类任务,随机梯度下降算法需要迭代 10K 次,并且每次迭代输入网络中的图像个数为 30.模型训练完成之后,根据第 2.2 节所描述的方法获取每幅图像的注意力图,并通过算法 2 生成伪像素标注.之后,利用伪像素标注来训练分割网络,其训练的学习率、动量及权值衰减等设定与分类任务相同.分割网络的训练迭代次数为 8 000 次.

用于获取注意力图及完成分割任务的网络训练时间分别约为 16h 和 10h.实验的配置为 12GB 显存的 NVIDIA GeForce TITIAN X.所有的实验均在深度学习 caffe 框架下完成.

3.2 注意力图及伪像素标注的有效性

正如第 2.2 节所述,本文可以产生具有类别信息的注意力图,而这些注意力图可以有效地提供图像中不同语义类别对象的位置信息.为了对比本文方法与 DCSM[45]方法的注意力图的准确性,本实验设定 4 组阈值 th=

{0.2,0.3,0.4,0.5}.由于注意力图具有类别信息,对于类别 c 的注意力图,如果空间位置 n 的值大于等于 th,则标注该位置的像素类别为 c,否则为背景.最后,将这些像素标注与真实像素标注做 IoU 运算,并计算 21 个类别的平均 IoU 值(如表 1 所示,采用 PASCAL VOC 训练集数据).表 1 中,DCSM 方法(第 2 列)的结果均使用 Shimoda 等人[45]所提供的代码,并且表 1 的所有结果均没有加入 dense CRF 后续优化过程.通过表 1 可以说明,本文所提方法生成的注意力图在所有阈值设定下所生成的伪像素标注都要比 DCSM 方法的结果更加准确(表 1 中,第 3 列为本文方法的结果).

Table 1 Mean IoU value (%) of pseudo pixel-level annotations generated by attention maps on 21 categories, under different foreground thresholds

表 1 多个前景阈值设定下,根据注意力图生成的伪像素标注在 21 个类别上的平均 IoU 值(%)

阈值	DCSM[45]	本文方法
0.2	33.64	38.66
0.3	36.01	41.78
0.4	36.48	**42.25**
0.5	35.40	40.61

为了进一步证明注意力图可以为后续的分割网络提供有用的信息,第 2 个实验将生成的注意力图直接看作分割结果,并且计算平均 IoU 值来验证注意力图的确提供了对象的类别信息.此外,为了验证逐次擦除法可以检测出图像中存在的多个对象,本实验比较了不同擦除次数下获取的显著图与注意力图融合得到的伪像素标注的准确性.表 2 列出了上述实验的对比结果.

Table 2 Validation results of attention map and pseudo pixel-level annotations

表 2 注意力图和伪像素级标注的验证结果

方法	平均 IoU(%)
AM	40.53
WSS[27]_S_1_AM	47.33
WSS[27]_S_2_AM	48.05
WSS[26]_S_3_AM	45.97
DHSN[28]_S_1_AM	53.11
DHSN[28]_S_2_AM	55.14
DHSN[28]_S_3_AM	55.20

表 2 采用 PASCAL VOC 2012 验证集.PASCAL VOC 数据库提供了验证集的准确类别标签,因此表 2 的实验根据图像的类别标签来生成注意力图,并且通过算法 2 生成伪像素标签.从表 2 可以看出,本文所提出的注意力图获取方法可以准确地提供图像中对象的位置信息,并且可以区分出图像中不同类别的对象(表 2,方法“AM”).同时,采用逐次擦除法获取的显著图与注意力图融合,可以生成更加精确的伪像素标注.WSS[27]方法基于图像级标签训练模型,DHSN[28]基于无类别(class-agnostic)的前景标注训练神经网络.这两个模型关注图像的前景对象,不具有任何语义信息.从而将具有类别信息的注意力图与显著图融合,可以提高伪像素标注的准确性.通过 WSS 方法获取的显著图与注意力图融合方法(表 2,方法“WSS[27]_S_1_AM”)使平均 IoU 值提高了 7 个百分点.而随着逐次擦除法的不断擦除,除了可以检测出当前显著的区域外,还可以检测出更多的对象,从而提升像素标注的准确性.

从表 2 可以看出,WSS 方法的逐次擦除法在获取 S_2 之后,平均 IoU 值最高(48.05%,表 2,方法“WSS[27]_S_2_AM”);而随着擦除次数的增加,其平均 IoU 值有所下降(45.97%,表 2,方法“WSS[27]_S_3_AM”).这说明过多次的擦除使得背景图像被检测,到从而降低了准确度.DHSN 方法在获取 S_3 之后,平均 IoU 值达到 55.20%,只比获取 S_2 的平均 IoU 值多 0.06 个百分点.从而说明显著图 S_2 已经检测到图像中存在的大部分前景对象.

图 4 列出了图像基于上述方法获取的分割结果.

图 5 展示了 WSS 方法和 DHSN 方法经过不同次数擦除后获取的显著图,可以看出,逐次擦除法可以检测出图像中的更多前景对象.

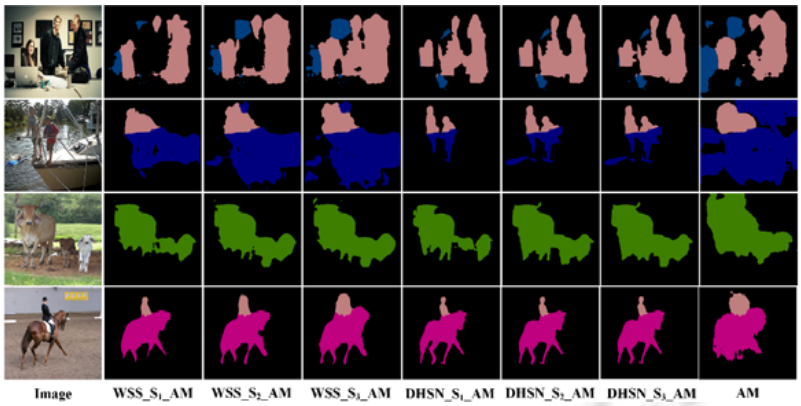

Fig.4 Visual segmentation results of the attention map and pseudo pixel-level annotations

图 4 注意力图及伪像素标注的可视化分割结果

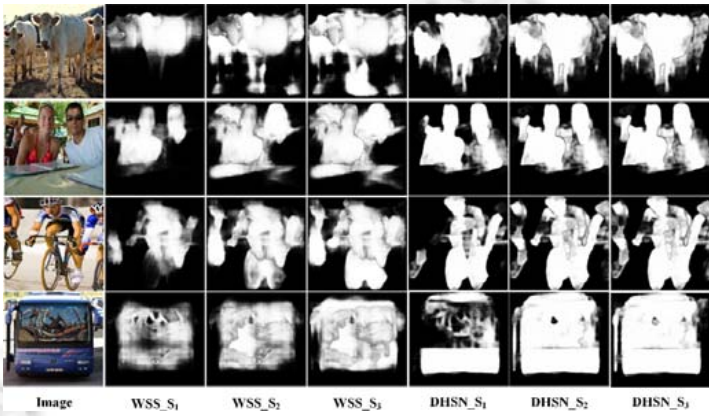

Fig.5 Saliency maps obtained by successive erasing method based on WSS and DHSN saliency network

图 5 基于 WSS 和 DHSN 显著网络的逐次擦除法获取的显著图

3.3 全连接条件随机场优化的有效性

为了验证条件随机场模型可以有效地检测到图像对象的轮廓从而提升分割的准确率,本次实验对第 3.2 节的实验结果通过 dense CRF 模型进一步优化.第 2.5 节已经给出了 dense CRF 的详细描述,由于第 3.2 节的实验结果指定了每个像素的标签($label(y_n)$),而非每个像素属于 21 个语义类别的概率,因此本实验将公式(6)的第一项$\phi(y_n)$改为

$$\phi(y_n=c)=\begin{cases}\lambda, & \text{if } c=label(y_n)\\ (1-\lambda)/|Z|, & \text{if } c\neq label(y_n) \text{ and } c\in\{Z,c_0\}\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

其中,$\lambda\in(0.5,1)$,表示 $y_n=c$ 的概率.实验中,我们设定$\lambda=\{0.7,0.8,0.9,0.98\}$.表 3 给出了在不同$\lambda$值的情况下 dense CRF 优化的分割结果.

从表 3 可以看出,随着λ增加,这两个方法的平均 IoU 值均呈上升趋势.这就说明当λ值增加时,我们给予通过融合显著图 S_2 与注意力图所生成的伪像素标注的置信度越高,从而通过 dense CRF 优化得到更准确的分割结果.对于方法"DHSN[28]_S_2_AM",当$\lambda=\{0.7,0.8,0.9\}$时,其平均 IoU 值均小于没有加入 dense CRF 优化过程的分割结果;特别是当λ=0.7 时,其平均 IoU 值为 51.45%,比无 dense CRF 优化过程的 55.14%要小近 4 个百分点.而当λ=0.98 时,其分割结果才有所提升(56.03%).这说明基于 DHSN 模型的逐次擦除法及融合算法 2 获取的伪像素标签比较准确,需要在 dense CRF 模型中给与较高的置信度(即属于该标签的概率更高),才能充分地发挥 dense

CRF 模型优化对象轮廓,提升分割准确率的优势.综上所述,dense CRF 模型可以优化其粗糙的分割结果,更加准确地分割出对象的轮廓.图 6 展示了当λ={0.7,0.98}时,方法"WSS[27]_S_2_AM"与"DHSN[28]_S_2_AM"通过 dense CRF 优化的分割结果.

Table 3 Segmentation results based on different λ values

表 3 基于不同λ值的分割结果

方法	Dense CRF	平均 IoU(%)
WSS[27]_S_2_AM	×	47.33
	λ=0.7	47.47
	λ=0.8	48.08
	λ=0.9	48.63
	λ=0.98	48.64
DHSN[28]_S_2_AM	×	55.14
	λ=0.7	51.45
	λ=0.8	52.65
	λ=0.9	54.11
	λ=0.98	56.03

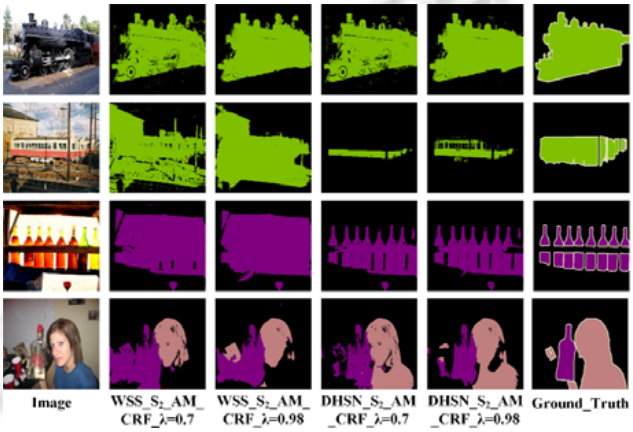

Fig.6 Comparison of visual segmentation results based on different λ values

图 6 基于不同λ值的可视化分割结果对比

3.4 分割网络

上述实验均验证了所获取的伪像素标注为分割网络提供有效的像素语义类别信息,从而引导分割网络的训练过程.本节实验比较了分别基于 WSS 与 DHSN 模型获取的显著图与注意力图融合生成伪像素标注(WSS_AM,DHSN_AM),并训练分割网络(DeepLab)的图像分割结果.表 4 给出了在 PASCAL VOC 2012 数据集上的实验结果,其中,第 3 列展示了训练集的伪像素标注与真实像素标注的平均 IoU 值,用于检测伪像素标注的准确率;第 4 例为加入分割网络之后在验证集上的分割结果.

Table 4 Performance comparison of segmentation network trained by pseudo pixel-level annotations based on different saliency maps

表 4 基于不同显著图生成的伪像素标注所训练的分割网络性能的比较

方法	显著图	训练集的平均 IoU(%)	验证集的平均 IoU(%)
WSS_AM_DeepLab	S_1	47.26	48.50
	S_2	48.80	53.30
	S_3	47.82	50.21
DHSN_AM_DeepLab	S_1	49.77	53.39
	S_2	53.14	54.88
	S_3	53.48	54.87
AM_DeepLab	×	42.25	41.39

从实验结果可以看出,单独使用注意力图作为像素标注并训练分割网络并不能取得分割性能的最大提升(表 4 中,方法“AM_DeepLab”).而通过融合显著图与注意力图生成的伪像素标注更加准确,从而为分割网络提供了更有效的信息,指导网络参数学习过程.从表 4 可以看出,最好的分割结果达到 54.88%,比最低的分割结果 41.39%提升了 13.5 个百分点.同时,在表 4 中我们也列出了不同擦除次数下获取的显著图对于最后分割结果的影响(表 4,第 2 列).从平均 IoU 变化趋势上可以看出,当显著图为 S_2 时(擦除两次),WSS_AM_DeepLab 和 DHSN_AM_DeepLab 均获得最优的分割效果(分别是 53.39%,54.88%).这个结论与表 2 所反映的实验情况基本相同.图 7 展示了上述实验方法的分割结果,其中最后 1 列为图像的真实分割.

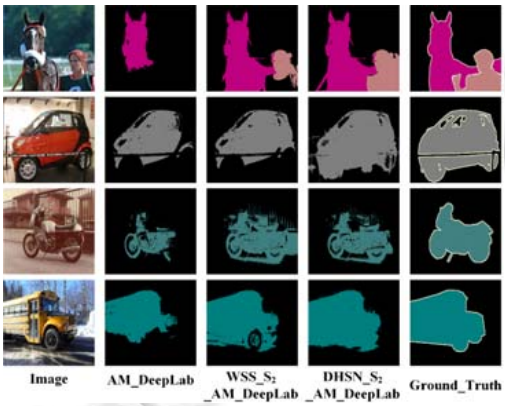

Fig.7 Comparison of segmentation results by using different methods to obtain the pseudo pixel-level annotations

图 7 使用不同方法获得伪像素级标注的分割结果对比

3.5 与弱监督图像分割方法的比较

本节列出了本文所提方法与其他弱监督方法在 PASCAL VOC 2012 验证集与测试集上的比较结果(表 5).从表 5 可以看出,当采用 DHSN[28]的显著图时,本文所提方法在验证集和测试集上分别获得 54.9%,55.3%的平均 IoU 值,其分割结果要明显优于除 AE-PSL[22]之外的其他方法.与 AE-PSL 方法相比,我们的结果仅仅下降了 0.1 与 0.4 个百分点.但是 AE-PSL[22]方法需要在每次获取对象位置信息之后擦除图像并重复训练分类网络,其迭代次数无法确定,而且分类网络与分割网络是两个独立的网络.而本文方法无需重复训练显著图模型,从而降低了模型训练时间,并且分类网络与分割网络共享网络结构(从模型训练时间来讲,假设 AE-PSL[22]方法需要迭代 3 次来训练分类网络,以 VGG-16 网络结构为参照,其分类网络的训练时间约为 16h×3=48h(GPU 配置:NVIDIA GeForce TITIAN X),其分割网络的训练时间约为 10h,总共约为 58h.而本文所提方法的模型训练时间约为 16h+10h=26h.由此可见,本文方法的模型训练时间要比 AE-PSL[22]方法降低了一半的时间).当采用 WSS[27]的显著图时,本文方法“WSS_S_2_AM_DeepLab”比 AE-PSL 方法下降了 1.7 个百分点,但是该方法是真正意义上的只采用图像级标注,而 AE-PSL 方法所使用的显著图[51]则需要更精确的前景标注.此外,一些对比方法使用了除图像类别标注的其他信息,例如,MIL[52]与 SN_B[39]均借助 MCG[53]模型产生对象分割候选集,从而提升分割结果;AugFeed-SS[54]使用 selective search[55]获取对象的分割候选集;STC[41]使用额外的图像数据集(50K flickr)与 PASCAL VOC 数据集共同训练分割网络.因此可以说明,本文方法通过最简单的计算框架得到更好的分割结果.

另外,从实验结果可以看出,我们的方法“DHSN_S_2_AM_CRF(λ=0.98)”在验证集上得到最优的分割结果(56.0%),但是在测试集上,其平均 IoU 值下降了近 7 个百分点.这是因为在获取注意力图时使用了验证集的类别标签,即明确地给出图像类别的注意力图,从而融合显著图获取伪像素标注.而在测试集合上,图像的类别标签是不可知的,因此首先需要通过分类网络判别图像的类别,并根据这些类别生成注意力图.可见分类网络是存在误差的,所以其分割准确性要比方法“DHSN_S_2_AM_DeepLab”降低很多.这同时也间接地说明训练分割网络的必要性,单纯的分类网络是不能够很好地提升分割准确率的.DCSM[45]仅通过分类网络获取注意力图,并用

dense CRF 优化分割结果,其分割准确率 45.1%明显低于本文的方法.

Table 5 Weakly supervised semantic segmentation results on validation and test images

表 5 在验证和测试数据集上的弱监督语义分割结果

方法	平均 IoU(验证)(%)	平均 IoU(测试)(%)
EM-Adapt[35]	38.2	39.6
LCEM-Fixed-2-Hyb[43]	45.4	46.0
CCNN[19]	36.3	35.6
MIL[52]	42.0	40.6
SN_B[39]	41.9	43.2
SEC[16]	50.7	51.7
STC[41]	49.8	51.2
Combining Cues[56]	52.8	53.7
DCSM[45]	44.1	45.1
AugFeed-SS[54]	52.6	52.7
Two-phase[44]	53.1	53.8
AE-PSL[22]	55.0	55.7
Build on FG/BG[40]	46.6	48.0
DHSN_S_2_AM_CRF(λ=0.98) (ours)	**56.0**	**49.7**
WSS_S_2_AM_DeepLab (ours)	**53.3**	**53.9**
DHSN_S_2_AM_DeepLab (ours)	**54.9**	**55.3**

表 6 列出与全监督和半监督语义分割方法的比较结果.DeepLab[2]的训练集中有 10 582 幅图像的像素级标注,因此该方法的平均 IoU 值在 65%以上.Bbox-EM-Fixed[35]与 BoxSup[34]方法均借助了 bounding box 标注准确地定位到对象的位置,虽然 bounding box 标注并不能提供对象的轮廓信息,但是却极大地降低了模型的训练难度,因此可以将使用这类标注的方法视为半监督方法.可见,这两个方法的平均 IoU 值较高.ScribbleSup[33]方法的训练集有 scribble 标注,该标注能够勾勒出对象的流型走向.What's the point[14]与 TransferNet[38]虽然分别采用了点标注和其他数据集的像素标注,但是这两个方法的分割性能均比本文方法要低.这说明仅仅依赖点标注来学习模型的参数是远远不够的,因为它并不能提供有效的对象位置信息.此外,由于数据集之间存在一定的差异,因此借助于其他数据集的像素标注来做目标数据集的分割任务存在一定的弊端.从该实验可以看出,本文所提方法与其他方法相比,在仅有图像级标注时仍然具有很好的分割性能.本文方法虽然没有比半监督或者全监督方法的平均 IoU 值高,但是在弱监督的设定下,提升了分割效果的同时也缩小了两者的差距.

Table 6 Comparison results with fully supervised and semi supervised approaches on PASCAL VOC 2012 segmentation dataset

表 6 与全监督、半监督方法在 PASCAL VOC 2012 分割数据集上的比较结果

方法	平均 IoU 验证集&测试集	监督方式|标注方法
DeepLab[2]	67.6% & 70.3%	全监督训练
Bbox-EM-Fixed[35]	64.8% & 69.0%	半监督训练,bounding box 标注
ScribbleSup[33]	71.3% & 73.1%	scribble 标注
What's the point[14]	42.7% & 43.6%	点(point)标注
BoxSup[34]	62.0% & 64.2%	bounding box 标注
TransferNet[38]	52.1% & 51.2%	MSCOCO 数据集像素标注
DHSNet_S_2_AM_CRF (λ=0.98) (ours)	56.0% & 49.7%	弱监督
WSS_S_2_AM_DeepLab (ours)	53.3% & 53.9%	弱监督
DHSN_S_2_AM_DeepLab (ours)	54.9% & 55.3%	弱监督

此外,表 7 列出了本文方法在 PASCAL VOC 2012 验证及测试集合上 21 个类别详细的分割结果.其中,分割最好的 5 个类别是"car、bird、cat、bus、airplane".可以看出,这 5 个类别中,"bus"和"airplane"通常占据图像的大部分区域,并且背景比较简单,轮廓单一,具有固定的形状;类别"car"与"bird"虽然不会占据图像的大部分区域,但是其出现的场景较为单一,例如路边、公路上、枝头、水面上等;此外,类别"cat"虽然出现的场景比较多样化,但是与背景的区分度较高,因此分割情况较为容易.分割效果较差的 5 个类别为"chair、diningtable、sofa、plant、bike".类别"bike"与"chair"都具有较复杂的轮廓,并不具有固定的形状;而"diningtable"通常与"chair"同时出现,经常会出现将"chair"分割为"diningtable"的情况;类别"sofa"通常与背景的区分度较低,出现的场景较为复杂,而

类别“plant”占据图像的区域较小,常常被放置在角落,并且轮廓复杂,因此也同样不容易被分割出来.

Table 7 Our segmentation results for each class on validation and test sets

表 7 本文方法在验证和测试集上针对每个类别的分割结果

类别	WSS_S_2_AM_DeepLab 验证集(%)	DHSN_S_2_AM_DeepLab 验证集(%)	WSS_S_2_AM_DeepLab 测试集(%)	DHSN_S_2_AM_DeepLab 测试集(%)
background	85.0	86.8	85.8	87.2
airplane	69.0	73.7	68.7	74.3
bike	25.5	26.3	28.9	30.9
bird	67.9	67.6	67.6	71.2
boat	49.6	57.2	39.3	44.8
bottle	62.1	65.9	57.5	61.6
bus	72.9	73.6	70.5	72.2
car	61.3	66.6	61.4	64.1
cat	70.6	70.9	66.5	70.5
chair	18.2	13.5	21.2	16.7
cow	57.4	63.4	57.2	58.2
diningtable	32.1	14.9	34.5	22.8
dog	60.4	63.2	65.0	67.1
horse	57.4	59.3	59.7	59.2
motorbike	60.9	63.3	67.4	68.6
person	45.1	59.2	47.3	60.0
plant	30.5	34.0	39.5	37.4
sheep	65.7	65.0	65.9	68.7
sofa	29.0	20.5	34.7	22.1
train	59.4	59.3	52.3	57.1
tv/monitor	39.8	48.1	41.9	46.2
Average	53.3	54.9	53.9	55.3

图 8 展示了本文所提方法的分割结果.我们可以看出,方法“DHSN_S_2_AM_DeepLab”的分割结果更加准确.

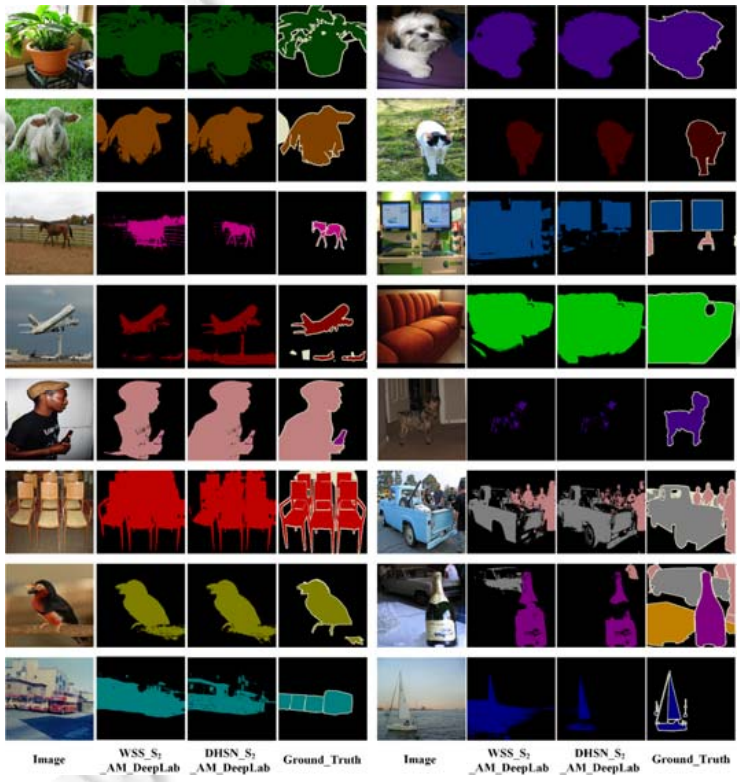

Fig.8 Visual segmentation results of the proposed methods on PASCAL VOC 2012 validation set

图 8 本文方法在 PASCAL VOC 2012 验证集上的可视化分割结果

4 结束语

本文提出一种基于图像级标注挖掘对象位置线索的弱监督图像分割方法.本文利用分类与分割共享的卷积神经网络生成具有类别信息的注意力图,该注意力图能够挖掘出对象的判别性区域.同时,本文采用逐次擦除法获取显著图,用于弥补注意力图丢失的对象空间位置信息,从而通过融合这两类信息生成伪像素标注并训练分割网络模型.通过实验可以说明,有效的融合注意力图与显著图可以提高伪像素标注的质量,并且间接地提升了弱监督分割的性能.通过在 PASCAL VOC 2012 数据集上与目前最先进的方法进行一系列的对比实验与分析,我们发现,本文所提的方法具有较好的分割准确率.

弱监督图像语义分割具有很好的应用前景.未来的工作将针对注意力图和显著图做进一步改进,希望通过图像的类别标签可以挖掘出更多的对象语义信息,进一步调整计算框架,并尝试应用于医学图像、遥感图像等新的领域.

References:

[1] Jiang F, Gu Q, Hao HZ, Li N, Guo YW, Chen DX, Survey on content-based image segmentation methods. Ruan Jian Xue Bao/ Journal of Software, 2017,28(1):160−183 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5136.htm [doi: 10.13328/j.cnki.jos.005136]

[2] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. of the IEEE, 1998, 86(11):2278−2324. [doi: 10.1109/5.726791]

[3] Krizhevsky A, Sutskever I, Hinton GE. ImageNet: Classification with deep convolutional neural networks. In: Proc. of the Advances in Neural Information Processing Systems. 2012. 1097−1105.

[4] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the Int'l Conf. on Learning Representation. 2015.

[5] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 770−778.

[6] Huang G, Liu Z, van der Maate L, Weinberger KQ. Densely connected convolutional networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 4700−4708.

[7] Bai C, Huang L, Chen JN, Pan X, Chen SY. Optimization of deep convolutional neural network for large scale image classification. Ruan Jian Xue Bao/Journal of Software, 2018,29(4):1029−1038 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5404.htm [doi: 10.13328/j.cnki.jos.005404]

[8] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 3431−3440.

[9] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected CRFS. In: Proc. of the Int'l Conf. on Learning Representation. 2015.

[10] Lin G, Milan A, Shen C, Reid I. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 1925−1934.

[11] Hariharan B, Arbelaez P, Girshick R, Malik J. Hypercolumns for object segmentation and fine-grained localization. In: Proc. of the IEEE Conf. on Computer Vision andPattern Recognition. 2015. 447−456.

[12] Mostajabi M, Yadollahpour P, Shakhnarovich G. Feedforward semantic segmentation with zoom-out features. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 3376−3385.

[13] Hariharan B, Arbelaez P, Girshick R, Malik J. Simultaneous detection and segmentation. In: Proc. of the European Conf. on Computer Vision. 2014. 297−312.

[14] Bearman A, Russakovsky O, Ferrari V, Li FF. What's the point: Semantic segmentation with point supervision. In: Proc. of the European Conf. on Computer Vision. 2016. 549−565.

[15] Xu J, Schwing AG, Urtasun R. Learning to segment under various forms of weak supervision. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 3781−3790.

[16] Kolesnikov A, Lampert CH. Seed, expand and constrain: three principles for weaklysupervised image segmentation. In: Proc. of the European Conf. on Computer Vision. 2016. 695−711.

[17] Vasconcelos M, Vasconcelos N, Carneiro G. Weakly supervised top-down image segmentation. In: Proc. of the IEEE Conf. on Computer Vision and PatternRecognition. 2006. 1001−1006.

[18] Xie WX, Peng YX, Xiao JG. Weakly-supervised image parsing via constructing semantic graphs and hypergraphs. In: Proc. of the Int'l Conf. on Multimedia. 2014. 277−286.

[19] Pathak D, Krahenbuhl P, Darrell T. Constrained convolutional neural networks for weakly supervised segmentation. In: Proc. of the Int'l Conf. on Computer Vision. 2015. 1796−1804.

[20] Papadopoulos DP, Clarke ADF, Keller F, Ferrari V. Training object class detectors from eye tracking data. In: Proc. of the European Conf. on Computer Vision. 2014. 361−376.

[21] Oh SJ, Benenson R, Khoreva A, Akata Z, Frittz M, Schiele B. Exploiting saliency for object segmentation from image level labels. In: Proc. of the IEEE Conf. on Computer Vision and PatternRecognition. 2017. 4410−4419.

[22] Wei YC, Feng JS, Liang XD, Cheng MM, Zhao Y, Yan SC. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 6488−6496.

[23] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 2921−2929.

[24] Zhang J, Lin Z, Brandt J, Shen X, Sclaroff S. Top-down neural attention by excitation backprop. In: Proc. of the European Conf. on Computer Vision. 2016. 543−559.

[25] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Proc. of the Int'l Conf. on Learning Representations. 2014.

[26] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-Cam: Visual explanations from deep networks via gradient-based localization. In: Proc. of the Int'l Conf. on Computer Vision. 2017. 618−626.

[27] Wang LJ, Lu HC, Wang YF, Feng MY, Wang D, Yin BC, Ruan X. Learning to detect salient objects with image-level supervision. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 136−145.

[28] Liu N, Han JW. Dhsnet: Deep hierarchical saliency network for salientobject detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 678−686.

[29] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: Proc. of the Int'l Conf. on Computer Vision. 2015. 1520−1528.

[30] Shen T, Lin G, Shen C, Reid I. Learning multi-level region consistency with dense multi-label networks for semantic segmentation. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence. 2017. [doi: 10.24963/ijcai.2017/377]

[31] Yu CQ, Wang JB, Peng C, Gao CX, Yu G, Sang N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proc. of the European Conf. on Computer Vision. 2018. 325−341.

[32] Zhao HS, Shi JP, Qi XJ, Wang XG, Jia JY. Pyramid scene parsing network. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 6230−6239.

[33] Lin D, Dai JF, Jia JY, He KM, Sun J. Scribblesup: Scribblesupervised convolutional networks for semantic segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 3159−3167.

[34] Dai JF, He KM, Sun J. Boxsup: Exploiting bounding boxes to supervise convolutiona networks for semantic. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 1635−1643.

[35] Papandreou G, Chen LG, Murphy K, Yuille AL. Weakly- and semisupervised learning of a deep convolutional network for semantic image segmentation. In: Proc. of the IEEE Conf. on Computer Vision. 2015. 1742−1750.

[36] Pathak D, Shelhamer E, Long J, Darrell T. Fully convolutional multi-class multipleInstance learning. In: Proc. of the Int'l Conf. on Learning Representation. 2015.

[37] Cinbis RG, Verbeek J, Schmid C. Weakly supervised object localization with multi-fold multiple instance learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015,39(1):189−203.

[38] Hong S, Oh J, Han B, Lee H. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 3204−3212.

[39] Wei YC, Liang XD, Chen YP, Jie ZQ, Xiao YH, Zhao Y, Yan SC. Learning to segment with image-level annotations. Pattern Recognition, 2016,59:234−244.

[40] Saleh F, Akbarian MSA, Salzmann M, Petersson L, Gould S, Alvarez JM. Built-in foreground/background prior for weakly-supervised semantic segmentation. In: Proc. of the European Conf. on Computer Vision. 2016. 413−432.

[41] Wei YC, Liang XD, Chen YP, Shen XH, Cheng MM, Zhao Y, Yan SC. STC: A simple to complex framework for weakly-supervised semantic segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017,39(11):2314−2320.

[42] Shen T, Lin G, Liu L, Shen C, Reid I. Weakly supervised semantic segmentation based on co-segmentation. arXiv:1705.09052, 2017.

[43] Li Y, Liu Y, Liu G, Zhai DM, Guo MZ. Weakly supervised semantic segmentation based on EM algorithm with localization clues. Neurocomputing, 2018,275:2574−2587.

[44] Kim D, Cho D, Yoo D, Kweon IS. Two-phase learning for weakly supervised object localization. In: Proc. of the Int'l Conf. on Computer Vision. 2017. 3554−3563.

[45] Shimoda W, Yanai K. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In: Proc. of the European Conf. on Computer Vision. 2016. 218−234.

[46] Hou QB, Dokania PK, Massiceti D, Wei YC, Cheng MM, Torr P. Bottom-up top-down cues for weakly-supervised semantic segmentation. In: Proc. of the Int'l Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition. 2017. 263−277.

[47] Oquab M, Bottou L, Laptev I, Sivic J. Is object localization for free? Weakly supervised learning with convolutional neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 685−694.

[48] Krahenbuhl P, Koltun V. Efficient inference in fully connected crfs with Gaussian edge potentials. In: Proc. of the Advances in Neural Information Processing Systems. 2011. 109−117.

[49] Everingham M, Eslami SMA, Gool LV, Williams CKI, Winn J, Zisserman A. The pascal visual object classes challenge: A retrospective. Int'l Journal of Computer Vision, 2015,111(1):98−136.

[50] Hariharan B, Arbelaez P, Bourdev L, Maji S, Malik J. Semantic contours from inverse detectors. In: Proc. of the IEEE Conf. on Computer Vision. 2011. 991−998.

[51] Jiang HZ, Wang JD, Yuan ZJ, Wu Y, Zheng NN, Li SP. Saliency object detection: A discriminative regional feature intergration approach. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2013. 2083−2090.

[52] Pinheiro PO, Collobert R. From image-level to pixel-level labeling with convolutional networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 1713−1721.

[53] Arbelaez P, Pont-Tuset J, Barron JT, Marques F, Malik J. Multiscale combinatorial grouping. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 328−335.

[54] Qi XJ, Liu ZZ, Shi JP, Zhao HS, Jia JY. Augmented feedback in semantic segmentation under image level supervision. In: Proc. of the European Conf. on Computer Vision. 2016. 90−105.

[55] Uijlings JRR, Van De Sande KEA, Gevers T, Smeulders AWM. Selective search for object recognition. Int'l Journal of Computer Vision, 2013,104(2):154−171.

[56] Roy A, Todorovic S. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 3529−3538.

附中文参考文献:

[1] 姜枫,顾庆,郝慧珍,李娜,郭延文,陈道蓄.基于内容的图像分割方法综述.软件学报,2017,28(1):160−183. http://www.jos.org.cn/1000-9825/5136.htm [doi: 10.13328/j.cnki.jos.005136]

[7] 白琮,黄玲,陈佳楠,潘翔,陈胜勇.面向大规模图像分类的深度卷积神经网络优化.软件学报,2018,29(4):1029−1038. http://www.jos.org.cn/1000-9825/5404.htm [doi: 10.13328/j.cnki.jos.005404]

李阳(1987－),女,博士,主要研究领域为机器学习,图像处理,计算机视觉.

刘国军(1979－),男,博士,副教授,CCF 专业会员,主要研究领域为机器学习,计算机视觉,模式识别.

刘扬(1976－),男,博士,副教授,CCF 专业会员,主要研究领域为机器学习,图像处理,计算机视觉.

郭茂祖(1966－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器学习,生物信息学.